What are Outliers?

- **Outlier** → a data object that deviates significantly from the rest of the objects
  - *Ex: a student with exceptionally high grades*
- Normal versus anomalous data objects → *how to define normalcy?*
- Outliers versus noise →*randomness, repetition patterns*

Global	Contextual	Collective
• Deviate significantly from the rest of the dataset • Ex: *Intrusion Detection* • How to measure deviation?	• Deviate *with respect to context* (time, location) • Ex: *Temperature values* • *Contextual* attributes used to evaluate context • *Behavioral* attributes used to evaluate outlier behavior	• A *subset* of objects collectively deviates significantly from the dataset • Ex: *Multiple order delays, DoS attacks*

Global: بتبقي outliers بعيده عن الداتا

Context: لازم احدد attribute معينه وهي اللي بتحدد context زي time, location مثلا بتبقي حاجه غريبه في وقت معينه يكون فيه بيع كتير لمنتج معين في وقت الشراء

Collective: لو حصل حركات بيع وشاء عادي بس لو حصل بيع في وقت اكتر من الشراء بيبقي collective، ازاي احدد الباكيت اللي جايه كويسه ولا باكيت هاتسبب هاك عليه و ازاي اتعامل مع noise

Challenges for Outlier Detection

Modeling normal objects & outliers
- Normal models are challenging to build
- Distinction between normalcy and anomaly is ambiguous

Application-specific outlier detection
- How much deviation is considered an anomaly/outlier?

Handling noise in outlier detection
- How to detect outliers in the presence of noise?
- Noise sometimes "hides" outliers!

Outliers Detection Methods

Statistical Methods

- **Parametric**

يقدر يحدد outlier بطريقتين χ^2 , boxplots ودول درسناهم في المحاضرات اللي قبل كده

  - For **univariate outliers** → *boxplots* → parameters are *mean* and *IQR*
  - For multivariate outliers → χ^2-*statistic* → parameter is *mean*

- **Nonparametric** → learn normal model from input data

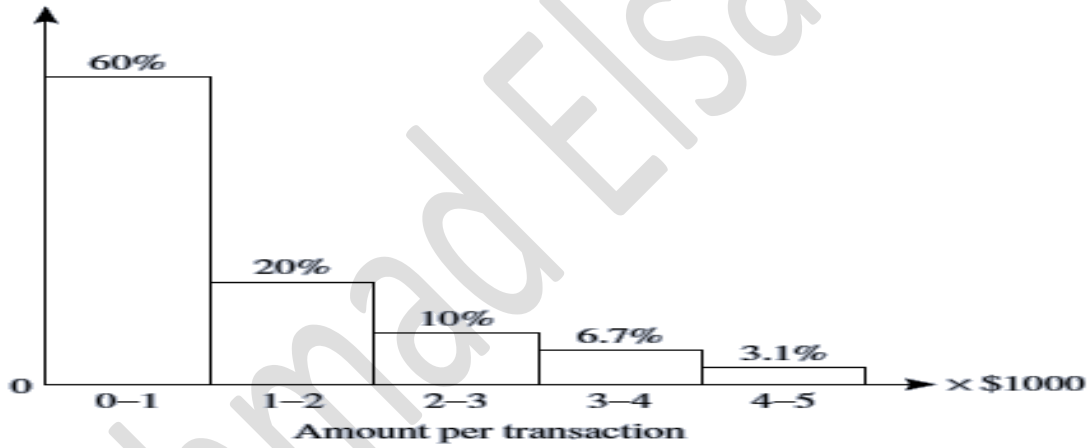
او اننا بنشتغل علي output data ودي عن طريق histogram

  - *histograms*

A transaction with the amount of $7500 is considered an outlier

- Does not belong to any of the bins (0.2% of transactions > $5000)

بيحسب الداتا لو لقي قيمه ملهاش frequency هايعتبرها outlier



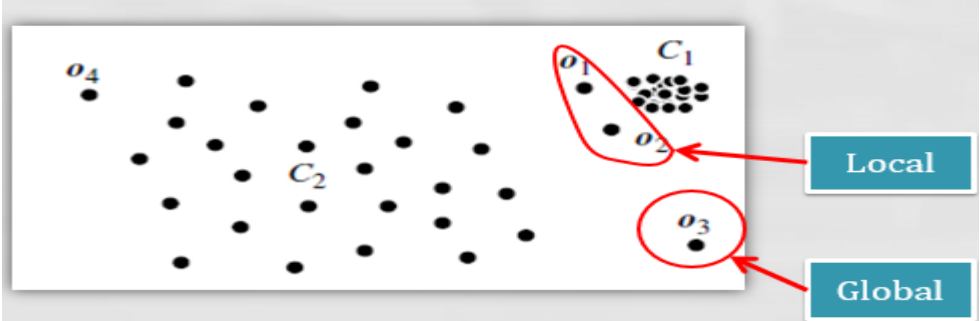
Proximity-based Methods

- **Distance-based** → for an object ***o***, examine the number of other objects in its ***r-neighborhood***
  - ***r*** is a distance threshold
  - π is a fraction threshold → min # objects needed in neighbourhood

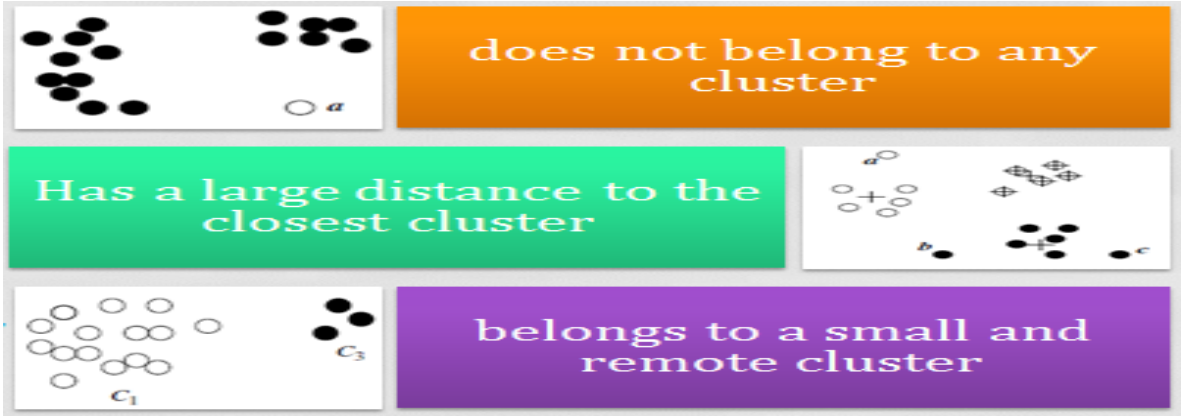
هامسك كل نقطه في دائره معينه وشوف عدد النقط اللي جواها لازم يكون جواها عدد معين عشان اقول عنهم انهم مش outlier (يعني اقل عدد موجود مع بعض عشان اقول انهم مش outlier)

- **Density-based** → for an object o, examine its density *relative to* the density of its local neighborsA local outlier factor (LOF) is computed in terms of the *K-NN* of an object in comparison to its neighbours

هاشوف اذا كان كل الداتا outlier ولا لا وبحسب local outlier factor وبحسبه عن طريق KNN



Clustering-based Methods



ممكن احدد outlier عن طريق clustering

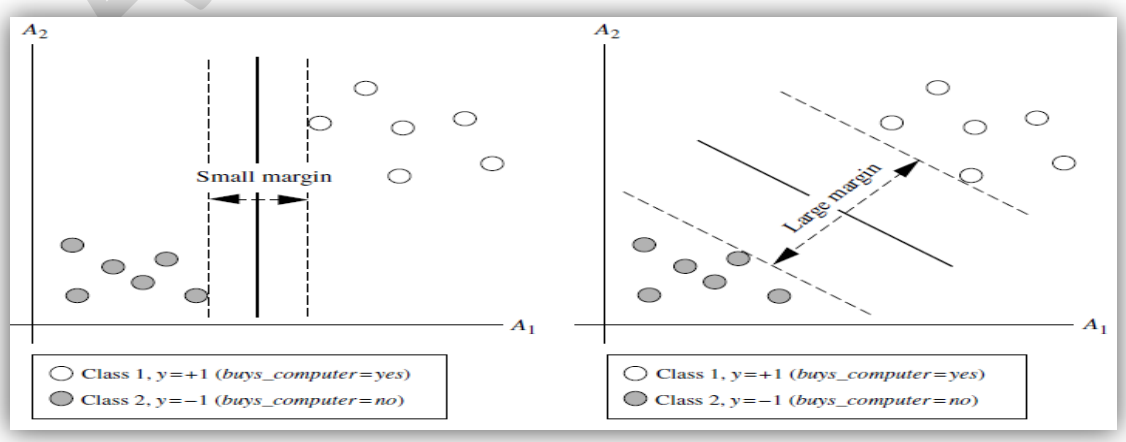
يا اما الاقي نقطه مابتنتميش لاي كلستر او المسافه بينها واقرب كلستر كبيرهه او بتنتمي لكلستر صغير وبيعيد عن بقيت الكلستر التانين

Classification-based Methods

- Use training data to build a normal model (one class)
  - Any *samples that do not belong to normal class* are outliers

بستخدم training data عشان ابني كلاس اي داتا هلاقيها مش تبع الكلاس ده يبقي outlier

- Learn the **decision boundaries** of the normal class
  - Using SVMs for example (this was not discussed in class)



Mining Contextual and Collective Outliers

- **Contextual outliers**:
    1. Determine contextual attributes

       احدد attribute اللي هاستخدمها في contextual
    2. Group those attributes' values

       اجمع قيم attribute
    3. Determine context of object (its group)

       احدد كل جروب
    4. Do conventional outlier detection <u>within that group</u>

       احدد outlier داخل كل جروب
    5. Ex → customers with similar *age* who live within *same area* may have *similar behaviour* → a customer in that age group and living in that area with a different behaviour is an outlier
- **Collective outliers** → challenging and advanced area